Natural genetic variation caused by small insertions and deletions in the human genome

Ryan E. Mills,[1,8] W. Stephen Pittard,[2] Julienne M. Mullaney,[3,4] Umar Farooq,[3] Todd H. Creasy,[3] Anup A. Mahurkar,[3] David M. Kemeza,[3] Daniel S. Strassler,[3] Chris P. Ponting,[5] Caleb Webber,[5] and Scott E. Devine[1,3,4,6,7,9]

[1]*Department of Biochemistry, Emory University School of Medicine, Atlanta, Georgia 30322, USA;* [2]*Bimcore, Emory University, Atlanta, Georgia 30322, USA;* [3]*Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA;* [4]*Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA;* [5]*MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom;* [6]*Winship Cancer Institute, Emory University, Atlanta, Georgia 30322, USA;* [7]*Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA*

Human genetic variation is expected to play a central role in personalized medicine. Yet only a fraction of the natural genetic variation that is harbored by humans has been discovered to date. Here we report almost 2 million small insertions and deletions (INDELs) that range from 1 bp to 10,000 bp in length in the genomes of 79 diverse humans. These variants include 819,363 small INDELs that map to human genes. Small INDELs frequently were found in the coding exons of these genes, and several lines of evidence indicate that such variation is a major determinant of human biological diversity. Microarray-based genotyping experiments revealed several interesting observations regarding the population genetics of small INDEL variation. For example, we found that many of our INDELs had high levels of linkage disequilibrium (LD) with both HapMap SNPs and with high-scoring SNPs from genome-wide association studies. Overall, our study indicates that small INDEL variation is likely to be a key factor underlying inherited traits and diseases in humans.

[Supplemental material is available for this article. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE27889. The INDEL variants reported in this study have been deposited in the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) (a complete listing of the accession numbers can be found in Supplemental Table 17).]

The age of personalized genomics is well under way. Human genomes are being sequenced at unprecedented rates (Levy et al. 2007; Bentley et al. 2008; Ley et al. 2008; Wang et al. 2008; Wheeler et al. 2008; Ahn et al. 2009; Kim et al. 2009; McKernan et al. 2009; The 1000 Genomes Project Consortium 2010; Schuster et al. 2010), and projects are under way to sequence the genomes of at least 1000 additional humans (Hayden 2008; The 1000 Genomes Project Consortium 2010). The long-range goal of these studies is to "crack the code" of natural genetic variation, i.e., to understand how changes in our genetic blueprints influence human traits. The prime motivation for these studies is to improve human health by understanding how genetic variation affects health, diseases, and medical treatments. Personal genome sequences will be used to predict the future health of individuals and to develop customized medical treatments that are optimized based on the genetic variation that is detected.

A critical first step in this process is to gain a comprehensive knowledge of the genetic variation that is harbored by human populations and to develop databases of informative variants that can be used to predict human health. Several types of natural genetic variation have been identified in humans, including single nucleotide polymorphisms (SNPs) (The International SNP Map Working Group 2001; The International HapMap Consortium 2003, 2005; The 1000 Genomes Project Consortium 2010), small insertions and deletions (INDELs) ranging from 1 bp to 10 kb in length (Weber et al. 2002; Bhangale et al. 2005; Mills et al. 2006; Korbel et al. 2007; Kidd et al. 2008; The 1000 Genomes Project Consortium 2010), and larger structural variants ranging from 10 kb to several megabases in length (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005; Conrad et al. 2006, 2010a,b; Hinds et al. 2006; McCarroll et al. 2006; Kidd et al. 2008; The 1000 Genomes Project Consortium 2010; Mills et al. 2011). Minisatellites, microsatellites, and transposon insertions also have been identified within these variant classes (Weber et al. 2002; Mills et al. 2006, 2007). Although SNPs and larger structural variants have received considerable attention, small INDELs remain largely under-discovered, and methods for studying these INDELs have lagged behind methods for analyzing other forms of variation. As a consequence, we know relatively little about the impact of small INDEL variation on human biology and diseases.

In this study, we identified almost 2 million small INDELs in the genomes of 79 diverse humans. More than 800,000 of these small INDELs mapped to human genes, including the coding exons and promoters of these genes. Several lines of evidence indicate that coding INDELs in particular are likely to affect gene function in humans. We also developed new microarray-based INDEL genotyping technologies to study the population genetics of small INDELs in diverse humans. Overall, our study indicates that small INDEL variation is extensive in human genomes and is likely to have a major impact on human biology and diseases.

[8]Present address: Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA.
[9]Corresponding author.
E-mail sdevine@som.umaryland.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.115907.110.

21:830–839 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

Results

Small INDEL discovery using DNA sequencing data from 79 humans

To gain a better understanding of small INDEL variation in human populations, we examined 98 million Applied Biosystems (Sanger) DNA re-sequencing traces that had been deposited into the trace archive at the National Center for Biotechnology Information (NCBI) (Supplemental Table 1). Many of these traces were generated previously by genome centers for SNP discovery projects or for BAC sequencing projects that were not used in published genome assemblies. Traces that were generated from targeted, PCR-based re-sequencing projects were excluded from our analysis. The final trace set includes DNA sequence information from 79 diverse humans, making it an ideal resource for variation discovery (Supplemental Table 2). We confirmed that these traces provide excellent coverage of the human genome (Supplemental Fig. 1).

By comparing these 98 million traces to build hg18 of the reference human genome sequence (The International Human Genome Sequencing Consortium 2001), we identified 1.96 million nonredundant small INDELs from an initial discovery set of 3.4 million INDELs (Table 1; Supplemental Tables Chr1–ChrY; see Methods). We previously established that our INDEL discovery pipeline has a validation rate of 97.2% (Mills et al. 2006; please see Methods). We also determined that 1.36 million of the 1.96 million INDELs in this study (69.3%) were discovered in more than one independent trace or could be confirmed in the chimpanzee (The Chimpanzee Sequencing and Analysis Consortium 2005) or Celera (Venter et al. 2001) genomes (Table 1). Thus, additional validation was achieved through these comparisons.

Table 1. Summary of small INDELs identified from 98 million re-sequencing traces

Traces screened	98,350,511
Bases analyzed (trimmed)	42,000,606,219
Reference human genome (hg18)	3,107,677,273
Cumulative coverage	13.5×
Variants identified	3,400,787
Unique (nonredundant) variants	1,955,656
Insertions	978,721
Deletions	976,935
Total bases affected	11,909,159
Total double hit	1,355,228
Match chimp allele	878,574
Match celera allele	648,941
Match variant allele	688,990
Total double center	477,295
Transposon insertions (with TSDs)	1150
Alu	1004
L1-Ta/pre-Ta	70
SVA	58
INDELs overlapping SV size range	
>50 bp	7245
>1000 bp	957

98 million human traces were obtained from the NCBI Trace Archive and were compared to the reference human genome to identify potential variants. All of the data were assembled into a MySQL database and deposited into dbSNP (Supplemental Tables Chr1–ChrY). INDEL-positive traces were compared to (a) other INDEL-positive traces, (b) the chimp genome (The Chimpanzee Sequencing and Analysis Consortium 2005), and (c) the Celera genome (Venter et al. 2001) to determine whether the allele was identified in other genomes and could be classified as a "double hit" allele. INDEL-positive traces from independent projects and centers were tracked to provide independent confirmation (Supplemental Tables Chr1–ChrY). INDELs were mapped to the promoters, 5′ and 3′ UTRs, splice sites, and introns of genes. This and other annotations were tracked in a MySQL database (and are listed in Supplemental Tables Chr1–ChrY). INDELs that were caused by *Alu*, L1, and SVA retrotransposon insertions are listed in Supplemental Table 16.

Our INDELs were found on all 24 human chromosomes at an average spacing of one INDEL per 1589 bp of DNA. They ranged from 1 bp to 10,000 bp in length and followed a size distribution in which the majority of INDELs were <100 bp in length (Supplemental Fig. 2; Wheeler et al. 2008). Like known SNPs, which affect ~15 Mb of DNA (dbSNP build 129), our INDEL variants affected 11.9 Mb of the human genome. Thus, the amount of genetic variation that is caused by small INDELs, in terms of base pairs, is considerable and approaches that caused by known SNPs.

Comparisons with personal human genomes and populations

We next wished to compare our 1.96 million variants to the small INDELs that have been discovered in personal human genomes and populations. We first compared our variants to the INDELs that that have been deposited to dbSNP and found that 37% (726,871/1.96 million) of our INDELs had been deposited previously. Thus, 63% of our INDELs are novel compared to the INDELs in dbSNP (build 129). We also examined five of the personal human genomes that have been sequenced, including four "healthy" genomes (Levy et al. 2007; Bentley et al. 2008; Wang et al. 2008; Wheeler et al. 2008) and the genome of a patient with acute myelogenous leukemia (Ley et al. 2008). Twenty-two percent (432,958) of our 1.96 million INDELs were present in one or more of these genomes (Fig. 1; Supplemental Table 3). Finally, we compared our INDELs to the 1.48 million INDELs that recently were reported by the 1000 Genomes Project (Fig. 1C; The 1000 Genomes Project Consortium 2010). We determined that 463,377 of our 1.96 million INDELs (23.6%) were present in the 1000 Genomes Project data set (Fig. 1C). The relatively small overlap between our INDELs and the INDELs from these other studies (Fig. 1; Supplemental Table 3; dbSNP) suggests that INDEL discovery is likely to be incomplete in human populations.

Structural variants (SVs) and transposon insertions

Our INDEL discovery range (of 1 bp to 10,000 bp) overlaps the discovery ranges that have been defined for structural variants (SVs), copy number variants (CNVs), and transposon insertions. For example, the 1000 Genomes Project recently defined SVs as variants that are >50 bp in length (The 1000 Genomes Project Consortium 2010; Mills et al. 2011), and other groups have defined SVs as variants that are >1 kb (Iafrate et al. 2004). We identified 7245 variants that are >50 bp and 957 variants that are >1000 bp among our collection of 1.96 million INDELs (Table 1; Supplemental Table 4). Thus, a small fraction of our 1.96 million INDELs (0.4%) overlaps these larger variant classes. We compared these 7245 variants to: (1) SVs in the Database of Genomic Variants (Iafrate et al. 2004), (2) SVs that were reported recently by Conrad et al. (2010a,b), and (3) SVs that were reported recently by The 1000 Genomes Project Consortium (2010) (Mills et al. 2011). We found that 3582 of our 7245 variants (49.4%) were novel compared to these other variants (Supplemental Table 4). Many of the remaining 3663 variants provide breakpoint resolution for known SVs at the single-nucleotide level for the first time. We also suspected that some of our INDELs might have been caused by transposon insertions (Supplemental Fig. 2). Indeed, 1150 transposon

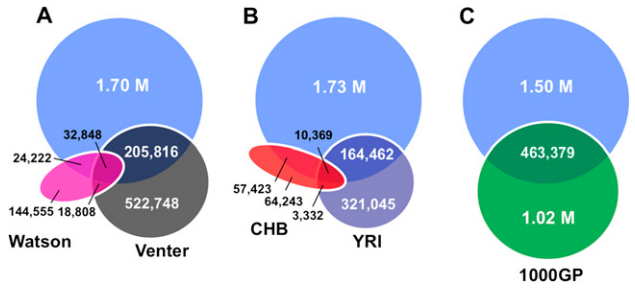


Figure 1. Comparisons of our data with small INDELs identified from other projects. (*A*,*B*) Diagrams comparing the 1.96 million INDELs discovered in this study with the small INDELs that were identified in four personal genomes. (*A*) Comparison of our 1.96 million INDELs (light blue, *top*) with Venter (Levy et al. 2007) and Watson (Wheeler et al. 2008) INDELs. (*B*) Comparison of our 1.96 million INDELs (light blue, *top*) with Han Chinese (Wang et al. 2008) and Yoruban (Bentley et al. 2008) INDELs, (*C*) Comparison of our 1.96 million INDELs (light blue, *top*) with the 1.48 million INDELs identified by the 1000 Genomes Project (1000GP) (The 1000 Genomes Project Consortium 2010).

insertions were identified in our INDEL collection, including *Alu*, L1, and SVA insertions (Table 1; Supplemental Table 16). A total of 170/1150 (14.8%) of these insertions previously were identified by our laboratory using transposon-seq methodologies (Iskow et al. 2010).

Small INDELs in human genes

Although INDELs could, in principle, affect a number of functional elements in the human genome, of particular interest are variants that alter human genes. A large number (819,363 or 42%) of the 1.96 million INDELs in this study mapped to known human genes, and 2123 INDELs (0.1%) affected the coding exons of these genes (Fig. 2; Supplemental Fig. 3; Supplemental Table 5). These "coding INDELs" were identified by comparing the coordinates of our INDELs to those of annotated RefSeq and Ensembl genes (build hg18). An initial set of 1715 coding INDELs was identified that affected 990 of the 20,705 RefSeq genes that were examined (equivalent to 4.8% of the RefSeq genes). An additional 408 coding INDELs were identified in 215 Ensembl genes that were not annotated in the RefSeq collection. Thus, a combined total of 1300 exons and 1205 genes were affected by coding INDELs in our collection (equivalent to ~5.8% of the genes in the human genome). These data indicate that apparently healthy humans harbor a substantial genetic load of coding INDELs.

A total of 184 of our 2123 (8.7%) coding INDELs overlapped one or more exon boundaries, and these variants generally would be expected to abolish gene function (Supplemental Fig. 3; Supplemental Table 5). The remaining coding INDELs (1939/2123 or 91.3%) fell entirely within coding exons. More than half of these INDELs (1037/1939 or 53.5%) would be expected to cause frameshifts that would introduce premature termination codons (PTCs). Some of these PTCs are predicted to occur at positions that would target the encoded mRNA for nonsense-mediated decay (NMD), and such variants generally would abolish gene function. However, PTCs that occur in the final 3′ exon (or in the 3′ 50 nt of the penultimate exon) may produce mRNAs that escape NMD; these mRNAs could, in principle, encode novel proteins that have gain-of-function or dominant-negative effects. The remaining coding INDELs (902/1939 or 46.5% of those within exons) were multiples of 3 nt, and thus, maintained the open reading frames (ORFs) of the original proteins (Supplemental Table 5).

Many (357/1205 or 29.6%) genes were affected independently by two or more coding INDELs (Supplemental Fig. 4; Supplemental Table 5). For example, nine in-frame exon variants of the *DSPP* gene (involved in the development of the dentin portion of teeth) (Mastrangelo et al. 2007) were identified (Supplemental Table 5). It is tempting to speculate that such polymorphisms might play a role in dental health. Moreover, 14 coding exon variants of the *EP400* tumor suppressor gene (Fuchs et al. 2001) were identified (Supplemental Table 5). A total of 52 genes that previously have been implicated in human cancers harbored coding INDELs (Supplemental Table 5). At least four gene families that previously were known to be highly polymorphic in human

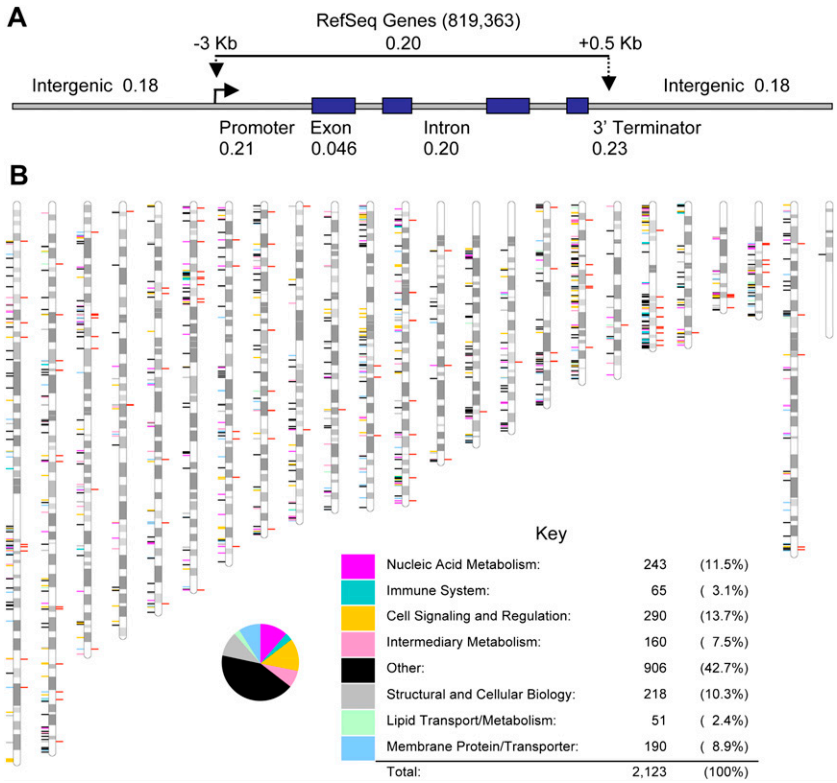


Figure 2. Distribution of coding exon variants in the human genome. (*A*) The figure depicts a typical RefSeq gene and its features. 819,363 small INDELs from our study were mapped to RefSeq genes. The INDEL-to-SNP ratios for each genomic compartment are indicated. (*B*) The 1205 genes that were affected by 2123 coding exon variants (Supplemental Table 5) are indicated on the map of human chromosomes (colored marks to the *left* of the chromosomes indicate an affected gene). Each mark is indexed by color to indicate gene function (and is cross-referenced to the pie chart *below*). A red mark to the *right* of each chromosome indicates that the affected gene previously was linked to a known disease. The pie chart shows the functional breakdown of the coding variants.

populations (Hughes and Yeager 1998; Shimomura et al. 2002; Suzuki et al. 2002; Gilad and Lancet 2003) also carried a large number of coding INDELs. In particular, 46 exon variants were identified in 23 different keratin genes; 14 exon variants were identified in seven collagen genes; 40 exon variants were identified in 32 olfactory receptor genes; and 11 exon variants were identified in two *HLA* genes (Supplemental Fig. 4; Supplemental Table 5). Thus, coding INDELs contribute to the high levels of genetic variation that are characteristic of these gene families. In each case, these high levels of genetic variation are known to be maintained for a biological purpose (Hughes and Yeager 1998; Shimomura et al. 2002; Suzuki et al. 2002; Gilad and Lancet 2003). These examples, along with the many additional examples in Supplemental Table 5, illustrate the point that coding INDELs may potentially cause a great deal of human biological diversity.

Coding INDELs that are likely to cause human phenotypes

We next set out to identify INDELs that are likely to cause phenotypic consequences in humans. Although some of our coding INDELs could, in principle, cause gain-of-function phenotypes, our analysis was focused solely on human INDELs that would disrupt gene function. To identify INDELs that are likely to be disruptive in humans, we identified INDELs that (1) reside within coding exons and cause frameshifts or overlap exon boundaries, (2) affect more than just the last exon, and (3) affect all Ensembl transcripts for the overlapped gene. Filtering on these three criteria resulted in 548 INDELs that affect 394 human genes (Supplemental Table 5). To gain insight into the likely phenotypic consequences of these disruptions, we identified 101 genes from this set that are associated with targeted disruption phenotypes in the mouse (Supplemental Table 6; Eppig et al. 2005, 2007). Of these 101 experimentally disrupted genes, 84 (83%) yield an abnormal phenotype upon homozygous disruption in the mouse (Supplemental Table 6).

We also identified five additional genes for which heterozygous disruptions are known to cause human diseases. As might be expected, the human disorders that are associated with these gene disruptions are nonlethal, often are late-onset, and are associated with relatively mild or variable phenotypes that might easily go undetected. The five genes, and their associated diseases, are *CEL* (maturity-onset diabetes of the young; OMIM: 609812) (Raeder et al. 2006), *IRAK3* (early onset asthma; OMIM: 611064) (Balaci et al. 2007), *KRT14* (ectodermal dysplasia syndromes; OMIM: 161000 and 125595) (Lugassy et al. 2006), *MYO6* (progressive hearing loss; OMIM: 606346) (Sanggaard et al. 2008; Hilgert et al. 2008), and *RP1* (retinitis pigmentosa; OMIM: 180100) (Supplemental Table 6; Guillonneau et al. 1999; Pierce et al. 1999; Sullivan et al. 1999). Three additional human genes with disruptive INDELs have mouse orthologs that are being used as human disease models in the hemizygous state. These are *CACNA1F* (congenital stationary night blindness; OMIM: 300071) (Mansergh et al. 2005), *NBN* (Nijmegen Breakage Syndrome; OMIM: 251260) (Dumon-Jones et al. 2003; Resnick et al. 2003; Tanzanella et al. 2003), and *TCOF1* (Treacher Collins syndrome; OMIM: 154500) (Supplemental Table 6; Dixon et al. 2000; Marzalek et al. 2003; Shoo et al. 2004). These analyses indicate that at least some of our INDELs are likely to have phenotypic consequences in humans, even when present in just one copy.

Evidence for strong purifying selection on coding INDELs

To investigate whether purifying selection might act on coding INDELs (particularly those that are detrimental), we next examined the relative distributions of INDELs compared to SNPs in the human genome. Our computational pipeline simultaneously detects both INDELs and SNPs from ABI trace data, and has high levels of accuracy with both classes of variation (Tsui et al. 2003; Mills et al. 2006; please see Methods). We determined that the genome-wide ratio of INDELs to SNPs was 0.19 (i.e., 1 INDEL for every 5.3 SNPs) (Supplemental Table 7). Both genes and intergenic regions had ratios that were similar to this genome-wide average (0.20 vs. 0.18, respectively) (Fig. 2A). Although most gene features (promoters, introns, terminators) also had similar ratios (Fig. 2A), the ratio for coding exons was well below this genome-wide average (0.046) (Fig. 2A). We confirmed that this difference was caused by a reduction of INDELs. This difference is most likely explained by the fact that INDELs create major changes in coding exons (including frameshifts and amino acid changes), whereas SNPs often produce synonymous changes that have little or no impact on gene function. Thus, strong purifying selection appears to eliminate INDELs that map to coding exons much more frequently than SNPs. Despite these reductions, coding INDELs were still fairly abundant in the genomes examined, indicating that coding INDELs are nevertheless likely to have a substantial impact on humans (see also below).

Microarrays for genotyping small INDELs in humans

Although microarrays have been used widely to genotype SNPs (The International HapMap Consortium 2003, 2005) and relatively large structural variants (Sebat et al. 2004; Conrad et al. 2006, 2010a; Hinds et al. 2006; McCarroll et al. 2006; Mills et al. 2011), microarray-based assays for interrogating small human INDELs have received little attention. This is particularly true for INDELs in the size range of 1 bp to 100 bp, which account for ~99% of the INDELs in our study. As a consequence, we lack robust tools to study our INDELs and the small INDELs that are being discovered in personal human genomes. Thus, we set out to develop new genotyping methods to interrogate small INDELs on microarrays. In particular, we developed assays to genotype INDELs that are 1 bp to 100 bp in length. We sampled several commercial platforms and found that the most successful approach was to adapt methods that originally were developed for the Affymetrix 5.0 and 6.0 SNP arrays (The International HapMap Consortium 2005). Like these arrays, we used reduced representation and statistical probe clustering models to develop assays for INDEL genotyping (Fig. 3). However, in contrast to Affymetrix SNP arrays, our probe sets were designed to specifically interrogate novel junctions that define small INDELs using a series of overlapping 25-mer probes (Supplemental Fig. 5). The probe design protocols, library files, and Affymetrix Power Tools (APT) software packages were modified accordingly to accommodate the unique format of our custom INDEL data. Overall, successful probe sets with good BRLLM-P clustering characteristics were developed for 10,003 INDELs using a confidence cutoff of 0.05 (Fig. 3D; Supplemental Table 8; Methods). Trace frequency data indicate that these 10,003 INDELs were representative of the larger INDEL collection (Supplemental Fig. 6). As a measure of accuracy, we also performed validation studies on a representative set of INDELs using PCR methodologies. The vast majority (268/271 or 99%) of the genotyping calls were identical for the two sets of measurements (Fig. 3D; Supplemental Table 9). Likewise, arrays that were hybridized with the same genomic DNA probes on separate days had >99% identical calls, indicating that our methods are highly reproducible (Supplemental Table 8).

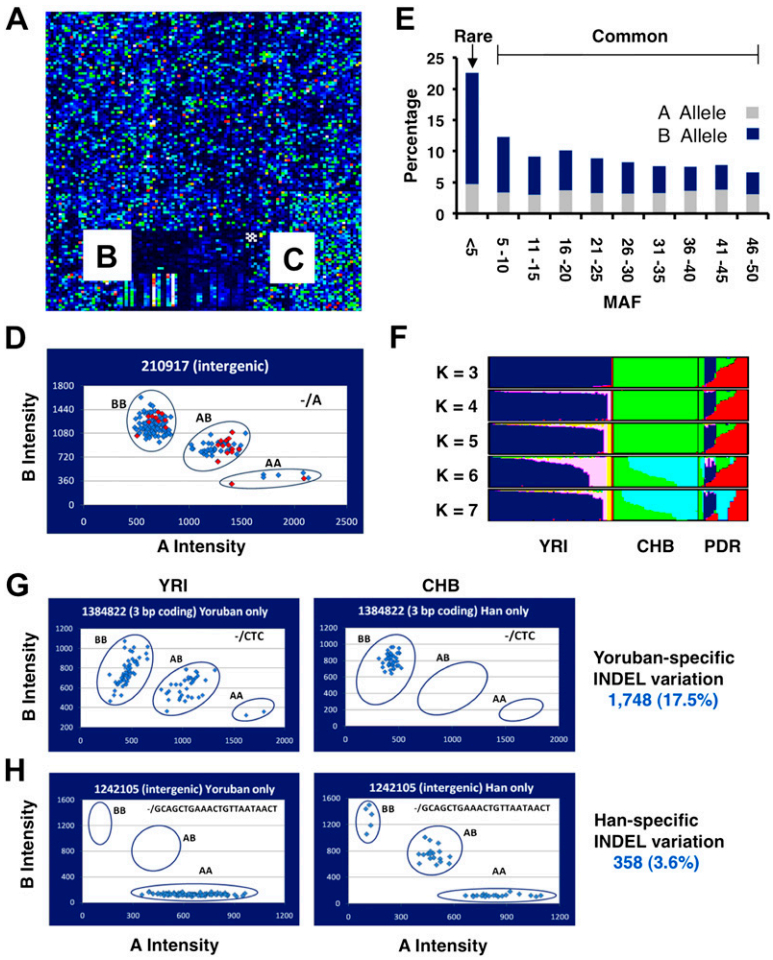


Figure 3. Affymetrix INDEL genotyping arrays. (*A–C*) A region of a custom Affymetrix INDEL microarray is shown following hybridization and scanning using protocols established for the Affymetrix 6.0 array. Section C contains 1500 Affymetrix SNPs that were developed for the HapMap project and are also present on the SNP 6.0 array. These were included as positive controls. The average cqc for our arrays after excluding arrays with scores below 0.4 was 2.2, with a range of 0.53–3.67. The call rate was 96.1%. Section B contains a manufacturing control. Section A represents the remainder of the array, which contains INDEL probes. (*D*) Plot of signal intensities for a typical set of INDEL probes following BRLLM-P analysis. Note that three distinct clusters were obtained for the three INDEL states (AA, AB, BB). PCR validation studies were conducted in parallel to evaluate the accuracy of the calls (Supplemental Table 9). A typical result is shown for INDEL 210917. The 24 individuals from the polymorphism discovery resource (PDR) (Collins et al. 1999) that were sampled by PCR are shown in red (the calls were 100% concordant between the arrays and the PCRs). The overall validation rate with 12 representative INDEL assays in 24 individuals was 99% (Supplemental Table 9). (*E*) Allelic frequencies. The allelic frequencies are plotted for the 10,003 INDELs that were examined on the INDEL microarrays (Supplemental Table 8). Although the majority of variation meets the definition of common genetic variation (where the minor allele has a frequency of ≥5%), rare INDELs also were identified. (*F*) Structure plot of INDEL data. The INDEL genotypes from our arrays were analyzed for population substructure. The PDR panel, which was designed to capture global diversity, has a large degree of substructure (as indicated by the colored peaks; *right*). The Yoruban (YRI) and CHB populations also have some residual substructure. (*G*) Population-specific INDEL variation. INDELs were identified where both INDEL alleles were present in one population but only one allele was present in the other. An example of a YRI-specific INDEL is shown. Note that both A and B alleles are present in the YRI population (and all three genotypes are present), whereas only the B allele (and one genotype) is present in the CHB population. The INDEL shown (1384822) is a 3-bp coding INDEL. (*H*) An example of a CHB-specific INDEL.

Population genetics of small INDELs

Overall, 8836 of the 10,003 INDELs on our arrays (88.4%) showed allelic variation in the panel of 158 diverse humans that was examined (Fig. 3E; Supplemental Table 8). For an additional 212 (2.1%) INDELs, we detected solely the trace (B) allele. Since the trace allele is highly abundant and also differs from the reference genome, these INDELs also were confirmed. This is a high confirmation rate (90.5%), given that we did not probe all of the original humans that were used for INDEL discovery. In many of the remaining cases (649/955 or 68%), the trace allele was confirmed in multiple traces, in the chimp genome, or in the Celera genome (or combinations of these three). Likewise, our INDEL detection pipeline has a validation rate of 97.2% (Mills et al. 2006), further supporting the conclusion that most of the remaining variants are likely to be rare but authentic INDELs.

Our microarrays allowed us to measure the allelic frequencies of our INDELs in the population that was examined (Fig. 3E; Supplemental Table 8). A total of 7728 (77.3%)of the 10,003 INDELs had minor allelic frequencies that were ≥5%, and the remaining 2275 (22.7%) had minor allelic frequencies of <5%. Thus, the majority of INDELs detected in our trace experiments meet the definition of common human genetic variation (Fig. 3E; Table 2; Supplemental Table 8). We also examined the allelic frequencies of INDELs that mapped to RefSeq genes (Table 2; Supplemental Tables 8, 10). Although only 22.7% of the 10,003 INDELs on the array were rare (MAF <5%), a much larger percentage of the coding INDELs on the array were rare (72/111 or 64.9%), suggesting that purifying selection was acting on these alleles. In contrast to the data in Figure 2A, these population frequency data indicate that even the coding INDELs that we can observe are enriched for deleterious variants compared to noncoding INDELs. A two-tailed Fisher's exact test revealed that this result was highly significant ($P = 6.6 \times 10^{-21}$) (Table 2). We also examined the number of triplet (in-frame) versus nontriplet (frameshifting) coding INDELs and observed an enrichment of rare alleles among the nontriplet (frameshifting) group. These data indicate that frameshifting INDELs are under the strongest negative selection ($P = 0.00036$), which is not surprising given that this class is expected to be the most disruptive. These data provide evidence that even the coding INDELs that have not yet been eliminated by selection are enriched for deleterious alleles and are likely to have an impact on humans.

We next examined INDEL inheritance patterns in the 30 Yoruban (YRI) trios that were examined in our study (containing a mother, father, and child). Ninety-nine percent of the INDELs

Table 2. Analysis of rare INDEL variants in RefSeq genes

Class	Number of rare/total	%	*P*-value
All INDELs on array	2,275/10,003	22.7	N/A
In RefSeq genes	1,050/4,466	23.5	0.32
Not in RefSeq genes	1,225/5,537	22.1	0.38
Exon (coding)	72/111	64.9	6.6×10^{-21}
Exon (noncoding)	32/122	26.2	0.39
Intron	895/4,044	22.1	0.44
Promoter	36/132	27.3	0.21
Terminator	15/57	26.3	0.53

For each class of INDELs examined, the number of rare INDELs (MAF <5%) versus the total number of INDELs in the class is listed. Two-tailed Fisher's exact tests were performed between the overall set of 10,003 INDELs that were genotyped on the array and each of the other classes that are listed. The data indicate that rare INDELs are over-represented in coding exons compared to the 10,003 INDELs. None of the remaining classes had such enrichment, suggesting that coding exons are the most sensitive to INDELs.

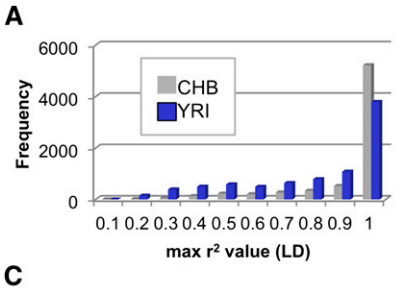
were inherited in a Mendelian fashion in the YRI trios, indicating that most of the small INDELs in our study are polymorphisms rather than de novo mutations (Supplemental Table 8). The remaining 1% of the INDELs had non-Mendelian transmission patterns. This class of INDELs is likely to be enriched for genotyping errors. It should be noted that since the error could be in the parent or the child, the per sample error rate is <1%. We also determined whether the allelic distributions were in agreement with Hardy-Weinberg predictions in each of the three subpopulations that were genotyped (i.e., the polymorphism discovery resource [PDR], YRI, and Han Chinese [CHB] populations) (Supplemental Table 11). An average of 96.8% of the INDELs were in Hardy-Weinberg equilibrium in the three populations (P = 0.01). Although statistical and genotyping errors alone can account for the non-HWE distributions observed, it is also possible that some of these unusual distributions were caused by sample relatedness (Fig. 3F) or natural selection.

We also compared the allelic frequencies of our INDELs in two diverse human populations (YRI and CHB) (Fig. 3F–H). A total of 2106 (21.1%) of the INDELs on our arrays had variation in one of the two populations but not in the other (Fig. 3G,H; Supplemental Table 12). Interestingly, most (1748) of these population-specific variants were YRI-specific variants (in these cases, the YRI population had both INDEL alleles but the CHB population had only one of the two alleles) (Fig. 3G). An additional 554 (5.5%) of the INDELs had allelic frequency differences of >0.5 between the two populations.

Linkage disequilibrium with SNPs

We next examined the linkage disequilibrium (LD) of our INDELs compared to the SNPs that have been genotyped in the YRI and CHB populations by the HapMap project (The International HapMap Consortium 2005). A total of 5218 INDELs had perfect LD (r^2 = 1) with at least one SNP in the CHB HapMap panel, and 6084 INDELs had useful LD (r^2 > 0.8) (Supplemental Table 13). Likewise, 5674 INDELs had useful LD with SNPs that were genotyped in the YRI HapMap panel (r^2 > 0.8). Most of the INDELs that had variation on our arrays (8836 or 88.4%) were in LD with at least one SNP in the CHB and/or YRI populations (Fig. 4A). These experiments demonstrate that thousands of small INDELs can be efficiently genotyped and integrated into the human HapMap with this approach.

Finally, we determined whether high-scoring SNPs from genome-wide association studies likewise had high levels of LD with any of our INDELs. Such associations could establish connections between our INDELs and human traits (including diseases). More than 56,000 SNPs from 118 published genome-wide association studies recently were collected into a single convenient database (Supplemental Table 14; Johnson and O'Donnell 2009). We used this database to determine whether any of the SNPs in these studies had high levels of pairwise LD with our INDELs. Indeed, this approach was highly successful, and 2290 SNPs were identified that had high levels of LD (r^2 > 0.8) with our INDELs (Supplemental Table 15). In fact, 1193 SNPs had perfect LD (r^2 = 1) with INDELs on our arrays (Fig. 4B). Almost half of the INDELs (1102 or 48%) were located in genes (Fig. 4B), and many of these INDELs could be envisioned to affect gene function (Fig. 4C). Thus, in conjunction



B

Class	Number
GWAS SNPs in LD with r^2 > 0.8	2,290
GWAS SNPs in perfect LD (r^2 = 1)	1,193
INDELs in genes (LD r^2 > 0.8)	1,102
exons total	62
coding INDELs	6
introns	991
promoters	29
3' terminators	20

C

SNP	Disease/trait	INDEL	r^2	Variant	Gene
INDELs in coding exons					
rs618145	Coronary artery	363513	1.0	-/GTGACAGCCTGCAAC	NM_015225
rs12685939	Type I diabetes	2235459	1.0	-/AAA	NM_153267
rs8067494	Type II diabetes	188688	1.0	-/CAA	NM_000676
rs8112054	2 hour glucose	1192262	1.0	-/ATA	NM_182515
rs7940239	Rheum. arthritis	1286116	0.84	-/GACA	NM_001004740
rs700760	Bone mass	1122616	0.83	-/TTT	NM_024617
INDELs in non-coding exons					
rs11259906	Crohn's disease	1010795	1.0	-/GT	NM_001717
rs1055055	Hypertension	253658	1.0	-/T	NM_173642
rs3790043	Parkinson's disease	1585720	1.0	-/C	NM_023112
rs2657879	Fasting glucose	1060257	1.0	-/AC	NM_207344
rs6791490	Triglyc/HDL ratio	2619148	1.0	-/AAC	NM_018010
INDELs in promoters					
rs6747513	ALS	361854	1.0	-/ATCATTATTACCACAGTG	NM_017750
rs11087256	Hypertension	2154277	1.0	-/AATA	NM_006363
rs4712969	Type I diabetes	816478	1.0	-/TGTAT	NM_001098486
rs1455453	Weight/Type 2 diab.	1813108	1.0	-/GCTACCTT	NM_021634
rs10507353	Crohn's disease	382264	1.0	-/A	NM_005977

Figure 4. Linkage disequilibrium between SNPs and INDELs. (*A*) The r^2 value was calculated for each SNP within a 1-Mb window of a given INDEL using the SNP genotypes that have been reported for HapMap 3 (http://hapmap.ncbi.nlm.nih.gov/) and our INDEL genotyping data from the same samples. For each population (YRI, CHB), the SNP with the maximum r^2 value was identified (Supplemental Table 13). INDELs in perfect LD with a SNP have an r^2 of 1.0. (*B*) LD also was examined for high-scoring SNPs (with *P*-values <0.001) that were identified in 118 GWAS studies (Supplemental Table 14; Johnson and O'Donnell 2009). GWAS SNPs that have high levels of LD (r^2 > 0.8) with INDELs are summarized. (*C*) Examples of INDELs from *B* that map to functional regions of genes. The 16 examples were taken from a larger collection of 1102 INDELs that have high levels of LD with GWAS SNPs and also map to genes (Supplemental Table 15).

with GWAS studies, our INDEL genotyping platform can be used to identify INDEL candidates that may influence human traits and diseases.

Discussion

We provide strong evidence that small INDEL variation is not only abundant in humans but is also likely to contribute to human phenotypic diversity. More than 800,000 of the small INDELs in our study mapped to human genes, including 2123 small INDELs that mapped to the coding exons of these genes (Fig. 2; Supplemental Tables Chr1–ChrY; Supplemental Table 5). Many of these coding INDELs would be expected to alter gene function (Fig. 2; Table 2; Supplemental Tables 5, 6; Taylor et al. 2004). We also identified more than 39,000 INDELs in the promoter regions of genes (Supplemental Tables Chr1–ChrY). Such INDELs could be envisioned to contribute to the allele-specific gene expression differences that have been observed in humans (Yan et al. 2002; Cheung et al. 2003; Cheung and Spielman 2009). Taken together, these two classes of INDELs alone (coding and promoter INDELs) are likely to harbor many variants that affect human biology.

Several independent lines of evidence indicate that coding INDELs in particular are likely to be detrimental in humans. First, we found that ~75% of the INDELs that arise in coding exons appear to be eliminated by strong purifying selection (Fig. 2A). Moreover, our INDEL genotyping experiments further indicated that most of the remaining coding INDELs are also under selective pressure as revealed by an over-representation of rare alleles in this class (Table 2). Taken together, these data indicate that coding INDELs in general, and frameshifting INDELs in particular, have negative effects on human gene function. A comparison of our most detrimental coding INDELs with targeted deletions of orthologous mouse genes provided additional support for this idea (Supplemental Tables 5, 6). Finally, many of our INDELs (including coding INDELs) were in perfect LD with high-scoring SNPs that have been identified previously in GWAS studies (Fig. 4). Therefore, it is likely that coding INDELs and any associated effects are being detected in these studies. Taken together, these observations suggest that coding INDELs (particularly those that cause frameshifts) are likely to be responsible for a great deal of phenotypic diversity and diseases in humans.

Coding INDELs also have been identified in at least some of the personal human genomes that have been sequenced. For example, several hundred coding INDELs have been identified in the Venter (Ng et al. 2008) and Watson (Wheeler et al. 2008) genomes. Although coding INDELs have not been reported in most of the other personal genomes that have been sequenced, it is not clear whether attempts were made to identify coding INDELs in these studies. Our data suggest that coding INDELs are likely to be present in all human genomes (Fig. 2; Supplemental Tables 5, 10). In further support of this conclusion, exome re-sequencing projects and the 1000 Genomes Project also have identified coding INDELs (Ng et al. 2009, 2010; The 1000 Genomes Project Consortium 2010).

INDEL microarrays for genotyping

Salathia et al. (2007) previously described an approach for genotyping INDELs in the size range of 25 bp to 7260 bp in *Arabidopsis thaliana*. These INDELs were genotyped using 70-mer oligonucleotide probes and Comparative Genome Hybridization (CGH). In contrast, our approach more closely resembles the methods that were developed for SNP genotyping on the Affymetrix 6.0 array (The International HapMap Consortium 2005). In particular, we derived statistical clustering models from population data to call the INDEL genotypes using BRLMM-P. However, in contrast to the Affymetrix 6.0 approach, we used a series of 25-mer probes that spanned the unique junctions of INDEL alleles (Supplemental Fig. 5). Thus, our approach is unique compared to both the CGH and the SNP-based approaches that are outlined above. The smaller (25-mer) oligonucleotide probes that were employed with our Affymetrix platform can discriminate alternative INDEL alleles that differ by as few as 1 to 25 bp, and these probes performed well over the entire range that was tested in our study (1 bp to 100 bp). This size range is ideal for human INDELs because >99% of human INDELs are smaller than 100 bp and most human INDELs are smaller than 25 bp (Supplemental Fig. 2). Although our initial array contained a relatively small number of INDEL assays (10,003), this approach is directly scalable to much larger Affymetrix array formats.

In the past, genotyping arrays have been extremely valuable for validating and genotyping the SNPs that have been discovered by DNA sequencing (The International HapMap Consortium 2005). Since INDEL array technologies have lagged behind SNP arrays, it has not been possible to perform similar follow-up studies on the small INDELs that have been discovered in personal human genomes and populations. Our approach now can be used to perform these studies. The several million INDELs that have been discovered in personal human genomes and populations now can be validated with our platform. This approach will be particularly useful for genotyping INDELs that have eluded detection with Illumina-based sequencing due to the complexities of mapping short reads that involve INDELs. Thus, even as the price of whole-genome sequencing continues to drop, array-based technologies are likely to play a complementary role in validating and genotyping INDELs.

Our study argues that small INDELs should be fully discovered, genotyped, and integrated into the human HapMap. INDELs that map to coding exons, promoters, and other functionally important sites could be given the highest priority. Our custom INDEL array technology now provides a means for accomplishing this goal. Larger array formats could be designed and manufactured to genotype the several million INDELs that have been discovered by us and others. Once integrated into the HapMap, these small INDELs could be readily detected through imputation, thus facilitating efforts to identify high-scoring markers and causative variants in GWAS studies. As efforts turn to routine whole-genome re-sequencing, small INDELs could be imputed from the SNPs that are detected using this integrated map of phased variation. Small INDELs can be technically challenging to detect with next-generation trace mapping, and this would provide an alternative way to detect and annotate small INDELs. The largest and most interesting challenge ahead will be to use medical histories, treatment successes, and other phenotypes to identify specific INDELs that affect human biology, with the overall goal of developing a comprehensive framework of predictive health.

Methods

INDEL discovery pipeline

Our INDEL discovery pipeline has been described previously (Mills et al. 2006). In this earlier study, we conducted a series of PCR-based validation studies with variants that were detected by our pipeline (Mills et al. 2006). In particular, we examined 215 variants

using PCR-based methodologies and validated 209/215 (97.2%) of the variants. Importantly, these validation studies were carried out with the same 24 humans that were used for INDEL discovery (24 humans from the polymorphism discovery resource) (Collins et al. 1999). Thus, we knew with certainty that any variant identified in these traces also should be found by PCR in the 24 humans. We examined a range of INDEL sizes and sequences in this validation study, including single-base-pair INDELs, repeat expansions, transposon insertions, coding INDELs, and SNPs. We achieved an overall validation rate of 97.2% (false discovery = 2.8%).

In our previous study, we found that a single ABI Sanger trace was sufficient to accurately identify INDELs >1 bp in length (from 2 bp to 10,000 bp) (Mills et al. 2006). However, single-base-pair INDELs were less accurately called by a single trace, and we implemented a double hit rule to call 1-bp INDELs. This led to a similarly high validation rate (97.3%) for single-base-pair INDELs (Mills et al. 2006). The number of times that a given INDEL was detected in traces is indicated in Supplemental Tables Chr1–ChrY. We also used the chimp and Celera genomes to evaluate whether the trace alleles could be validated by at least one additional source (Table 1; Supplemental Tables Chr1–ChrY; Mills et al. 2006).

Identification and analysis of INDELs from NCBI trace data

DNA traces were obtained from the trace archive at NCBI (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). Traces were processed and compared to the hg18 build of the human genome using quality scores to guide the analysis, as described previously (Mills et al. 2006). INDELs were defined as insertions or deletions in the 1-bp to 10,000-bp size range that could be detected by comparing ABI traces to the reference genome as outlined previously (Tsui et al. 2003; Bennett et al. 2004; Mills et al. 2006). All variants were entered into a Perl dbfile hash module to identify redundancies, and then into a MySQL database, which was used to store and analyze the data. Supplemental Tables Chr1 to ChrY contain all of our variants along with coordinates and other information. INDELs were mapped to RefSeq genes as follows. First, RefSeq gene tracks were obtained from the UCSC Human Genome Browser (http://www.genome.ucsc.edu). The coordinates of our variants were compared to the coordinates of all RefSeq genes to identify variants that overlapped these genes (and specific features within these genes). For coding INDELs, this process was repeated with the Ensembl gene track downloaded from the Ensembl site (http://www.ensembl.org), and a non-overlapping set was developed. Additional annotation was obtained from the RefSeq database at NCBI (http://www.ncbi.nlm.nih.gov/RefSeq) and OMIM (http://www.ncbi.nlm.nih.gov/Omim).

Affymetrix INDEL microarrays

Custom Perl scripts were written to identify INDELs that would be suitable for probing on Affymetrix arrays using protocols that have been developed for SNPs (http://www.affymetrix.com). For example, to adapt our INDELs to the reduced representation probing that is used for SNPs, we identified INDELs that fell within three size intervals of StyI/NspI restriction fragments: (1) 0.2–0.8 kb, (2) 0.8–1 kb, (3) 1 0–1.2 kb. Group 1 was the preferred size, but the other size ranges also were included, if necessary. Probes were developed as outlined in Supplemental Figure 5. Probes were compared to the RepeatMasker track of the human genome (build hg18) and were set aside if they overlapped known repeats. In some cases, probes were designed for INDELs involving repeat expansions and transposons, and these probes were allowed to contain repeats. Such probes generally did not perform well and were filtered out at later stages (see below). Probes also were compared to the SNPs and INDELs present in build 128 of dbSNP, and probes that overlapped SNPs or other INDELs were not used. One thousand five hundred SNP probe sets from the Affymetrix 6.0 SNP array were included on our arrays as positive controls and for quality-control analysis. Arrays were hybridized using the Affymetrix 6.0 array kits and protocols (http://www.affymetrix.com). The Affymetrix Power Tools (APT) software package (http://www.affymetrix.com/support/developer/powertools/changelog/index.html) was modified to handle INDEL cdf files and the multiple probe designs that are required for INDELs. CEL files were normalized, and the quality was assessed using the APT program apt-geno-qc. High-quality arrays (cqc >0.4) were analyzed with the APT program apt-probeset-genotype using BRLLM-P at a cutoff of 0.05.

The select probes feature in the Affymetrix APT software was used to identify the best probes from all of the initial probes that were included on the array. For each INDEL, between 12 and 24 probes (depending on the INDEL type) were included on the array to discriminate between the two INDEL states (Supplemental Fig. 5). The best-performing probes were identified using a combined cutoff of AIC value <325, and FLDAB >3. Both the AIC and FLDAB parameters are derived from the clustering data and provide measures of an assay's ability to discriminate the A and B states. The probes that were selected at these cutoffs were able to discriminate the two INDEL alleles very well and had excellent clustering properties in BRLLM-P as well as high validation rates. All other probes were filtered from further consideration. Assays that had more than 60 no calls also were removed. This approach led to a final set of 10,003 assays that were used for all remaining studies.

Acknowledgments

We thank Shari Corin and Paul Doetsch for critical reading of the manuscript. We thank Brian Cotton and Victor Felix for computational help and Stephanie Steiner for technical assistance with PCR experiments. We thank Andy Neuwald for advice on statistical analysis. We thank Mark Mazaitis for help with figures. We thank Brandon Young, Kurt Donner, Christofer Bertani, Eric Schell, Ali Pirani, and the Affymetrix Custom Array Program for technical assistance. We thank Jim Nemesh and Bob Handsaker for their help with assessing the LD properties of our genotyped INDEL set. We thank the centers that provided small INDEL data from personal genome sequences. This work was funded by grants from SUN Microsystems and the National Human Genome Research Institute, National Institutes of Health (F32HG004207 and R01HG002898).

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, et al. 2009. The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res* **19:** 1622–1629.

Balaci L, Spada MC, Olla N, Sole G, Loddo L, Anedda F, Naitza S, Zuncheddu MA, Maschio A, Altea D, et al. 2007. IRAK-M is involved in the pathogenesis of early-onset persistent asthma. *Am J Hum Genet* **80:** 1103–1114.

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168:** 933–951.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet* **14:** 59–69.

Cheung VG, Spielman RS. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* **10:** 595–604.
Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* **33:** 422–425.
The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.
Collins FS, Brooks LD, Chakravarti A. 1999. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8:** 1229–1231.
Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38:** 75–81.
Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010a. Origins and functional impact of copy number variation in the human genome. *Nature* **464:** 704–712.
Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME. 2010b. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* **42:** 385–391.
Dixon J, Brakebusch C, Fassler R, Dixon MJ. 2000. Increased levels of apoptosis in the prefusion neural folds underlie the craniofacial disorder, Treacher Collins syndrome. *Hum Mol Genet* **9:** 1473–1480.
Dumon-Jones V, Frappart PO, Tong WM, Sajithlal G, Hulla W, Schmid G, Herceg Z, Digweed M, Wang ZQ. 2003. Nbn heterozygosity renders mice susceptible to tumor formation and ionizing radiation-induced tumorigenesis. *Cancer Res* **63:** 7263–7269.
Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM, et al. 2005. The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res* **33:** D471–D475.
Eppig JT, Blake JA, Bult CJ, Richardson JE, Kadin JA, Ringwald M. 2007. Mouse genome informatics (MGI) resources for pathology and toxicology. *Toxicol Pathol* **35:** 456–457.
Fuchs M, Gerber J, Drapkin R, Sif S, Ikura T, Ogryzko V, Lane WS, Nakatani Y, Livingston DM. 2001. The p400 complex is an essential E1A transformation target. *Cell* **106:** 297–307.
Gilad Y, Lancet D. 2003. Population differences in the human functional olfactory repertoire. *Mol Biol Evol* **20:** 307–314.
Guillonneau X, Piriev NI, Danciger M, Kozak CA, Cideciyan AV, Jacobson SG, Farber DB. 1999. A nonsense mutation in a novel gene is associated with retinitis pigmentosa in a family linked to the RP1 locus. *Hum Mol Genet* **8:** 1541–1546.
Hayden EC. 2008. International genome project launched. *Nature* **451:** 378–379.
Hilgert N, Topsakal V, van Dinther J, Offeciers E, Van de Heyning P, Van Camp G. 2008. A splice-site mutation and overexpression of MYO6 cause a similar phenotype in two families with autosomal dominant hearing loss. *Eur J Hum Genet* **16:** 593–602.
Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38:** 82–85.
Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet* **32:** 415–435.
Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donohue PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949–951.
The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.
The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PW, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141:** 1253–1261.
Johnson AD, O'Donnell CJ. 2009. An open access database of genome-wide association studies. *BMC Med Genet* **10:** 6. doi: 10.1186/1471-2350-10-6.
Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Anonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64.
Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460:** 1011–1016.
Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.
Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254. doi: 10.1371/journal.pbio.0050254.
Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456:** 66–72.
Lugassy J, Itin P, Ishida-Yamamoto A, Holland K, Huson S, Geiger D, Hennies HC, Indelman M, Bercovich D, Uitto J, et al. 2006. Naegeli-Franceschetti-Jadassohn syndrome and dermatopathia pigmentosa reicularis: Two allelic ectodermal dysplasias caused by dominant mutations in DRT14. *Am J Hum Genet* **79:** 724–730.
Mansergh F, Orton NC, Vessey JP, Lalonde MR, Stell WK, Tremblay F, Barnes S, Rancourt DE, Bech-Hansen NT. 2005. Mutation of the calcium channel gene Cacna1f disrupts calcium signalling, synaptic transmission and cellular organization in mouse retina. *Hum Mol Genet* **14:** 3035–3046.
Marzalek B, Wisniewski SA, Wojcicki P, Kobus K, Trzeciak WH. 2003. Novel mutation in the 5′ splice site of exon 4 of the TCOF1 gene in the patent with Treacher Collins syndrome. *Am J Med Genet* **123A:** 169–171.
Mastrangelo F, Scioletti AP, Tranasi M, Tecco S, Sberna MT, Rinci R, Grilli A, Stuppia L, Gherlone E, Tete S, et al. 2007. Dentin sialophosphoprotein expression during human matrix development. *J Biol Regul Homeost Agents* **21:** 33–39.
McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38:** 86–92.
McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using tow base encoding. *Genome Res* **19:** 1527–1541.
Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16:** 1182–1190.
Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome?. *Trends Genet* **23:** 183–191.
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59–65.
Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic variation in an individual human exome. *PLoS Genet* **4:** e10000160. doi: 10.1371/journal.pgen.1000160.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461:** 272–276.
Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al. 2010. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat Genet* **42:** 790–793.
Pierce EA, Quinn T, Meehan T, McGee TL, Bersen EL, Dryja TP. 1999. Mutations in a gene encoding a new oxygen-regulated photoreceptor protein cause dominant retinitis pigmentosa. *Nat Genet* **22:** 248–254.
Raeder H, Johansson S, Holm PI, Haldorsen IS, Mas E, Sbarra V, Nermoen I, Eide SA, Grevle L, Bjorkhaug L, et al. 2006. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat Genet* **38:** 54–62.
Resnick IB, Kondratenko I, Pashanov E, Maschan AA, Karachunsky A, Togoev O, Timakov A, Polyakov A, Tverskaya S, Evgrafov O, et al. 2003. 657del5 mutation in the gene for Nijmegen breakage syndrome (NBS1) in a cohort of Russian children with lymphoid tissue malignancies and controls. *Am J Med Genet* **120:** 174–179.
Salathia N, Lee HN, Sangster TA, Morneau K, Landry CR, Schellenberg K, Behere S, Gunderson KL, Cavalieri D, Jander G, et al. 2007. Indel arrays: an affordable alternative for genotyping. *Plant J* **51:** 727–737.
Sanggaard KM, Kjaer KW, Eiberg H, Nurnberg G, Nurnberg P, Hoffman K, Jensen H, Sorum C, Rendtorff ND, Tranebjaerg L. 2008. A novel nonsense mutation in MYO6 is associated with progressive nonsyndromic hearing loss in a Danish DFNA22 family. *Am J Med Genet* **46:** 1017–1025.

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Peterson DC, Zhao F, Qi J, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463:** 943–947.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Shimomura Y, Aoki N, Schweizer J, Langbein L, Rogers MA, Winter H, Ito M. 2002. Polymorphisms in the human high sulfer hair keratin-associated protein 1, KAP1, gene family. *J Biol Chem* **277:** 45493–45501.

Shoo BA, McPherson E, Jabs EW. 2004. Mosaicism of a TCOF1 mutation in an individual clinically unaffected with Treacher Collins syndrome. *Am J Med Genet* **126A:** 84–88.

Sullivan LS, Heckenlively JR, Bowne SJ, Zuo J, Hide WA, Gal A, Denton M, Inglehearn CF, Blanton SH, Daiger SP. 1999. Mutations in a novel retina-specific gene cause autosomal dominant retinitis pigmentosa. *Nat Genet* **22:** 255–259.

Suzuki OT, Sertie AL, DerKaloustian VM, Kok F, Carpenter M, Murray J, Czeizel AE, Kliemann SE, Rosemberg S, Monteiro M, et al. 2002. Molecular analysis of collagen XVIII reveals novel mutations, presence of a third isoform, and possible genetic heterogeneity in Knobloch syndrome. *Am J Hum Genet* **71:** 1320–1329.

Tanzanella C, Antoccia A, Spadoni E, diMasi A, Pecile V, Demori E, Varon R, Barseglia GL, Tiepolo L, Maraschio P. 2003. Chromosome instability and nibrin protein variants in NBS heterozygotes. *Eur J Hum Genet* **11:** 297–303.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res* **14:** 555–566.

Tsui C, Coleman LE, Griffith JL, Bennett EA, Goodson SG, Scott JD, Pittard WS, Devine SE. 2003. Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res* **31:** 4910–4916.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **27:** 727–732.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Holt RA, Gocayne JD, Amanatides P, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456:** 60–65.

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. 2002. Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet* **71:** 854–862.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452:** 872–876.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. *Science* **297:** 1143.

Received September 27, 2010; accepted in revised form March 18, 2011.

Natural genetic variation caused by small insertions and deletions in the human genome

Ryan E. Mills, W. Stephen Pittard, Julienne M. Mullaney, et al.

Genome Res. 2011 21: 830-839 originally published online April 1, 2011
Access the most recent version at doi:10.1101/gr.115907.110